Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 7716–7720, October 1989
Biochemistry

Drosophila NK-homeobox genes

(NK-1, NK-2, NK-3, and NK-4 DNA clones/chromosome locations of genes)

YONGSOK KIM AND MARSHALL NIRENBERG

Laboratory of Biochemical Genetics, National Heart, Lung and Blood Institute, National Institutes of Health, Building 36, Room 1C-06, Bethesda, MD 20892

Contributed by Marshall Nirenberg, July 6, 1989

ABSTRACT Four *Drosophila melanogaster* homeobox genes were found by screening a genomic DNA library with oligodeoxynucleotides that correspond to a conserved amino acid sequence that is part of the putative site of homeobox proteins that recognizes nucleotide sequences in DNA. The amino acid sequences of NK-2, NK-3, and NK-4 homeoboxes are more closely related to one another (59–66% homology) than they are to other *Drosophila* homeoboxes (28–54% homology), whereas the homeobox of NK-1 is most closely related, in order of decreasing homology, to muscle segment homeobox, zerknüllt-1, NK-3, and distal-less homeoboxes. Three of the genes, NK-1, NK-3, and NK-4, comprise a cluster of homeobox genes located in the 93E1–5 region of the right arm of the third chromosome, whereas the fourth homeobox gene, NK-2, is located in the 1C1–5 region of the X chromosome.

Homeobox genes encode DNA binding proteins that regulate gene expression during development or in the adult (1–4). In most cases, the similarity between different kinds of homeobox proteins extends only over a segment of the protein that consists of 60–61 amino acid residues, the homeodomain, which is thought to be the portion of the protein that recognizes nucleotide sequences in DNA. Homeobox genes are particularly well expressed in nervous system, and the homeobox family of genes encodes the largest set of proteins that regulate gene expression in the nervous system that has been identified thus far (5–9).

In this report, we describe four newly discovered, related *Drosophila* homeobox genes that were detected with oligonucleotide probes corresponding to an amino acid sequence that is thought to be part of the nucleotide sequence recognition site of homeobox proteins.

METHODS AND MATERIALS

Oligodeoxynucleotides. An Applied Biosystems DNA synthesizer 380B was used to synthesize oligodeoxynucleotides. Oligonucleotides with the trityl groups attached were purified by OPC column chromatography and trityl groups then were removed as described by Applied Biosystems. [γ-^{32}P]ATP with a specific activity of 6000 Ci/mmol (1 Ci = 37 GBq) (New England Nuclear) was used for phosphorylation of oligodeoxynucleotides catalyzed by T4 polynucleotide kinase (10).

Detection and Cloning of Homeobox Genes. A *Drosophila melanogaster* genomic DNA library in Charon 4A (11) was obtained from the American Type Culture Collection. Recombinant phage [48,000 plaque-forming units (pfu)] and 2×10^9 *Escherichia coli* KH802 cells were plated in Petri dishes (150 mm) at a concentration of 12,000 pfu per dish. Four nitrocellulose replica filter plaque lifts were obtained from each Petri dish, and each filter was hybridized with a different [^{32}P]-oligodeoxynucleotide preparation [16–64 oligodeoxynucleotide species per preparation; 1.5×10^6 cpm/ml; 120–150 fmol/ml (the sum of all species of oligodeoxynucleotides)] at 37°C overnight and washed with a solution containing tetramethylammonium chloride at 53°C or 50°C for 30 min for 17-mers or 16-mers, respectively, as described by Wood *et al.* (12).

DNA Sequencing.* Cloned genomic DNA fragments cleaved by restriction enzymes were subcloned into Bluescript pKS+. Both strands of the homeobox regions of the following DNA fragments were sequenced by the dideoxynucleotide chain-termination method (13) using M13 universal primers or specific oligodeoxynucleotide primers and Sequenase 2 (United States Biochemical): NK-1, 1.4-kilobase (kb) *Eco*RI/*Pst* I DNA fragment; NK-2, 1.2-kb *Eco*RI/*Pst* I DNA fragment; NK-3, 0.7-kb *Pst* I DNA fragment; NK-4, 0.4-kb and 2.3-kb upstream *Hin*dIII DNA fragments. dITP was used to reduce compression of DNA bands.

Locations of Genes on Chromosomes. Salivary gland polytene chromosomes were hybridized with *Eco*RI-cleaved genomic DNA fragments that contained the appropriate homeobox region and had incorporated biotin-16 dUMP in place of some dTMP residues as described (14). Detek 1-HRP kits (Enzo Biochemicals) and the protocol supplied by the manufacturer were used.

RESULTS AND DISCUSSION

Detection of Homeobox Genes. The *Drosophila* genomic DNA library of Maniatis *et al.* (11) in Charon 4A was screened for recombinants corresponding to homeobox genes with five [^{32}P]oligodeoxynucleotide probe preparations designed to hybridize to highly conserved homeobox nucleotide sequences. The oligonucleotide preparations were 16 or 17 nucleotides long and each consisted of multiple species of oligodeoxynucleotides (described in the legend to Fig. 2). Replica filters were prepared and each filter was hybridized to a different [^{32}P]oligodeoxynucleotide preparation. The filters were washed under high-stringency conditions with a solution that contained tetramethylammonium chloride, which selectively binds to A·T base pairs and raises the melting temperature (t_m) of A·T base pairs to that of G·C base pairs (12). The t_m of each [^{32}P]oligodeoxynucleotide–DNA duplex then was dependent on the number of contiguous base pairs formed but was not affected by the proportion of G·C vs. A·T base pairs (12). Consequently, all species of 17-mer oligodeoxynucleotides hybridized to DNA were washed at the same temperature (53°C) for the stringent wash, and all 16-mers were washed at 50°C.

Of the 48,000 phage plaques that were screened, ≈200 clones were obtained that exhibited a positive autoradiographic signal with one of the five [^{32}P]oligonucleotide probe preparations, and 7 recombinant clones were obtained that gave positive signals with two or more probe preparations. Many of the 200 clones that were detected with only one

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*The sequences reported in this paper for NK-1 to NK-4 have been deposited in the GenBank data base (accession nos. M27289, M27290, M27291, M27292, respectively).

probe were cloned, but they were not studied further. The 7 clones that were detected with two or more ^{32}P-labeled probes were characterized by restriction site analysis and the nucleotide sequences of the homeobox regions of some of the clones were determined by using unlabeled probes as sequencing primers. Five of the 7 clones were found to be previously unknown homeobox genes. The 2 remaining clones correspond to known homeobox genes; 1 clone contains zerknüllt-1 (*zen-1*) and *zen-2* DNA (15), and the other clone corresponds to either *en* or *inv* (16) (data not shown).

Characterization of Homeobox Genes. In Fig. 1 are shown partial restriction site maps of the homeobox genes NK-1, NK-2, NK-3, and NK-4, the locations of homeobox regions within the DNA inserts, and the direction of transcription. The approximate chain lengths of the cloned NK-1 and NK-2 genomic DNA fragments are 15.0 and 14.1 kb, respectively. Three of the 7 clones detected with two or more probes correspond to NK-3. Clone 6 is a 14.7-kb DNA fragment that contains the NK-3 homeobox sequence. Clones 3 and 9 contain similar or identical DNA inserts (14.6 kb) that overlap clone 6 and contain both NK-3 and NK-4 homeobox sequences separated by ≈7.8 kb. The restriction site map of NK-3 and NK-4 shown is derived from data obtained from clones 6, 3, and 9. Subcloned *Eco*RI DNA fragments from clone 3 were used to determine NK-3 and NK-4 nucleotide sequences.

Partial Nucleotide Sequence of NK-1. The sequence of an 811-nucleotide portion of the NK-1 gene is shown in Fig. 2. The first 198 nucleotides correspond to the 3′ portion of an intron. Another intron, 217 nucleotides long, was found within the homeobox, between codons for homeobox amino acid residues 44 (glutamine) and 45 (valine). The intron–exon structure of the NK-1 gene was confirmed by sequencing NK-1 cDNA clones (to be described elsewhere). Three other *Drosophila* homeobox genes, labial (*lab*) (19, 20), abdominal-B (*Abd-B*) (21), and distal-less (*Dll*) (22) [Brista (23)] have introns at precisely the same location within the homeobox as NK-1. The intron within the NK-1 homeobox contains a nucleotide sequence for antennapedia (Antp) protein binding (18) and one or two binding sites for zeste protein [the consensus nucleotide sequence for zeste is TGAGYG (Y, pyrimidine) (17)].

The amino acid sequence of the initial portion of the first NK-1 exon shown in Fig. 2 is highly acidic—i.e., 12 of the first 26 amino acid residues shown are aspartyl or glutamyl residues. Twenty-five percent of the amino acid residues before the homeobox are glycine residues, which include 7 consecutive glycine residues, and 17% are serine or threonine residues.

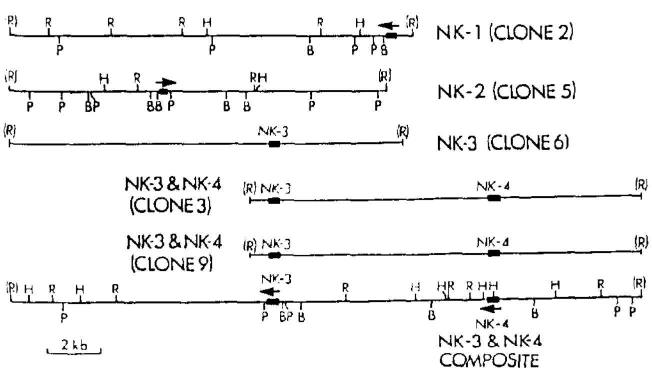

FIG. 1. Partial restriction maps of NK-1, NK-2, NK-3, and NK-4 cloned genomic DNA fragments. Solid boxes represent homeoboxes and are not drawn to scale. Arrows indicate direction of transcription. B, *Bam*HI; H, *Hin*dIII; P, *Pst* I; R, *Eco*RI; (R), *Eco*RI cloning site created by ligation of an *Eco*RI linker to genomic DNA.

Characterization of NK-2. The nucleotide sequence of the homeobox region of the NK-2 gene is shown in Fig. 3. The deduced amino acid sequence before the homeobox contains repetitive asparagine residues and a highly acidic region consisting of 14 aspartyl or glutamyl residues in a 31-amino acid segment (45% acidic amino acid residues). Twenty-five percent of the amino acid residues before the homeobox are glycine plus alanine. The carboxyl-terminal 30 amino acid residues of NK-2 are rich in histidine (20%), proline (17%), and glycine (17%). A 168-nucleotide 3′-untranslated region also is shown.

Characterization of NK-3 and NK-4. The nucleotide sequence and deduced amino acid sequence of part of the NK-3 homeobox gene are shown in Fig. 4. The initial part of the sequence consists of part of an exon that encodes 54 amino acids (17% alanine, 19% serine and threonine, and 9% asparagine), which is followed by a short, 119-nucleotide intron within the 26th codon before the homeobox. The intron–exon structure of the NK-3 gene was confirmed by sequencing NK-3 cDNA clones (K. Webber, Y.K., and M.N., unpublished data). The initial portion of the second

```
GAATTCATGGCACCAACATGTGCCGAAAAATTCCAATTAATCGAACAATGATGCGGTGG -293
(EcoRI)
CCGTGGTGATTGATTTCCGTTTTCCAATCCCCCAGGACATTGCCATTTGTCTGTGATGG -234

ATGGCCCTAGCCTGTTGACTTATGCAAAAAGAGAGACACCCGGAACTTATCGTGCCCAA -175

ATCTCCTCTTCTTTTTTTTGTCTTGCAG CC CAG GAT TTG AAT GAC ATG GAT -124
                                Gln Asp Leu Asn Asp Met Asp  -42

CAG GAC GAT ATG TGT GAC GAT GGC AGC GAT ATC GAC GAT CCC AGC  -79
Gln Asp Asp Met Cys Asp Asp Gly Ser Asp Ile Asp Asp Pro Ser  -27

AGC GAG ACG GAC TCC AAA AAG GGA GGC AGT CGT AAT GGG GAT GGA  -34
Ser Glu Thr Asp Ser Lys Lys Gly Gly Ser Arg Asn Gly Asp Gly  -12
                                         -1  +1
AAG TCC GGA GGT GGC GGC GGA GGT GGT TCA AAG CCT CGA CGA GCC   12
Lys Ser Gly Gly Gly Gly Gly Gly Gly Ser Lys Pro Arg Arg Ala    4

CGC ACC GCC TTC ACG TAC GAA CAA CTA GTT TCC CTG GAG AAC AAG   57
Arg Thr Ala Phe Thr Tyr Glu Gln Leu Val Ser Leu Glu Asn Lys   19

TTC AAG ACC ACC AGA TAT CTC AGC GTC TGC GAG CGA CTG AAC TTG  102
Phe Lys Thr Thr Arg Tyr Leu Ser Val Cys Glu Arg Leu Asn Leu   34

GCC CTC AGC TTG AGC CTG ACA GAG ACG CAG GTGAGCAATGATATATACT  151
Ala Leu Ser Leu Ser Leu Thr Glu Thr Gln                       44

CTATTGTTAAAGATTAAAATCCAGAGAAGTTATGTATATTTTGCAAAAAGTTGGTATAA  210

GTATTCTCTATGCTTTTCAATTTTAATAGAAGTAATTGAGTTAAAATATATTTTACTTT  269

GAGTGACTAAATTGAAAAGAAGTTCATTACTGTTTTTGAAATATTTAAATACCAATGTC  328

ATTTCTCATCATCCTTTTACAG GTT AAA ATT TGG TTC CAG AAC CGC CGC  377
                       Val Lys Ile Trp Phe Gln Asn Arg Arg   53

ACC AAG TGG AAG AAG CAG AAC CCC GGC ATG GAT GTC AAC TCC CCC  421
Thr Lys Trp Lys Lys Gln Asn Pro Gly Met Asp Val Asn Ser Pro   68

ACC ATC CCC CCG CCC GGC GGC GGC TCC TTC GGA CCG GG           460
Thr Ile Pro Pro Pro Gly Gly Gly Ser Phe Gly Pro Gly           81
```

FIG. 2. Nucleotide sequence and deduced amino acid sequence of the homeobox region of the NK-1 gene. Deoxynucleotide and amino acid residues are numbered on the right; 1 corresponds to the first deoxynucleotide or amino acid residue in the homeobox, which is enclosed in a large box. The acidic amino acid region is indicated by boxed Asp or Glu residues. Repetitive Gly residues before the homeobox also are enclosed in a box. The 1st and 2nd boxed nucleotide sequences in intron 2 are possible sites for binding of zeste protein to DNA (17). The 3rd site, ANNNNCATTA, is an Antp protein binding site (18). Arrowheads represent intron–exon junctions. Nucleotide 12 in an NK-1 genomic DNA clone was C, whereas the corresponding nucleotide residue found in an NK-1 cDNA clone was T. All oligodeoxynucleotide probes are complementary to the DNA strand shown; probe sequences, starting from the 5′-terminal nucleotide residues are as follows: —, probe 121, 24 species of 17-mers, (-)TTYTGRAACCA(T/A/G)TARAA; – –, probe 125, 48 species of 17-mers, (-)AACCA(T/G/A)ATYTTNACYTG; ---, probe 126, 64 species of 17-mers, (-)AAYTCYTTYTCNAGYTC; - -, probe 127, 64 species of 17-mers, (-)-C(T/G)RTTYTCRTTRAAYTC; · · ·, probe 130, 16 species of 16-mers, (-)A(C/G)T(C/G)(C/G)T(T/G)CTCCAGCTC. Y, pyrimidine; R, purine.

exon before the homeobox consists of 35% serine and 19% proline residues. The 54 amino acid residues after the homeobox are rich in alanyl and glycyl residues (22%) as well as leucyl residues (11%).

In Fig. 5 is shown the nucleotide sequence of part of the NK-4 gene. The initial part of the sequence consists of part of an intron, which is followed by an exon that contains the homeobox domain. The carboxyl-terminal region of the deduced NK-4 protein contains repetitive glutamine residues (M or opa repeats and a CAX repeat in the corresponding DNA).

Locations of Genes on Chromosomes. The cytological locations of the NK-1, NK-2, NK-3, and NK-4 genes in *Drosophila* third-instar larvae salivary gland polytene chromosomes are shown in Fig. 6. Unexpectedly, NK-1, NK-3, and NK-4 genes were found to reside in neighboring chromosomal bands in the right arm of chromosome 3. The NK-3 and NK-4 genes reside at 93E1–3, and the NK-1 gene resides at 93E3–5. When two probes, one for NK-1 and one for NK-3, were added to the same *in situ* hybridization reaction mixture, two labeled chromosomal bands were obtained at 93E1–5 that were separated only slightly. However, the NK-2 gene resides in the 1C1–5 region of the X chromosome. In Fig. 6*B*, the relative positions of the NK-3/NK-4 and NK-1 genes are shown correlated with the chromosomal bands in the 93E region of chromosome 3 in Bridges' revised map of chromosomal bands (24). These results show that NK-1, NK-3, and NK-4 comprise a cluster of homeobox genes.

Either NK-1, NK-3, or NK-4 genes may be the same as torso-like, a maternal effect gene that resides at 93E and is one of the ensemble of genes that determine the anterior–posterior pattern of the embryo (2). The torso-like gene and four other genes are required for the formation of both the anterior and posterior terminal, unsegmented portions of the embryo (the acron and telson) (2).

Another candidate is paired gene 9, which is thought to contain repetitive alternating codons for histidine and proline termed a paired repeat [also found in paired (25) and bicoid (25) genes] and a homeobox, which resides at 93E1–2 and was cloned by Frigerio *et al.* (25). Elsewhere, we will show that NK-1 contains a paired repeat (unpublished data); however, it is not known whether NK-1 is the same as paired gene 9. It should be noted that binding sites for polycomb protein have been detected at 93E1–4 (27). One of several candidates for the NK-2 gene is twisted, discovered by Demerec *et al.* (28), which is located between 1C-5 and 2C-10. The abdomens of adult *tw* mutants, viewed from behind, are rotated ≈30°C clockwise. However, further work is needed for the identification of NK-1, NK-2, NK-3, and NK-4 genes.

Homeobox Homology. The deduced amino acid sequences of the homeobox domains of NK-1, NK-2, NK-3, and NK-4 are shown in Fig. 7 and are compared with the 23 *Drosophila* homeobox sequences that have been reported thus far. NK-2, NK-3, and NK-4 homeoboxes are more closely related to one another (59–66% homology) than they are to other *Drosophila* homeoboxes. The maximum homology to a previously reported homeobox is to muscle segment homeobox (*msh*) (29) (54% homology). In order of decreasing homology, the homeobox of NK-1 is most closely related to *msh*, *zen-1*, NK-3 and *Dll* homeoboxes. NK-1, *lab*, *Dll*, and *Abd-B* genes may have originated from a common precursor because each gene contains an intron between the codons for the 44th and 45th homeobox amino acid residues. It has been suggested that homeobox proteins bind to DNA via a helix-turn-helix motif in the homeodomain and that amino acid residues 42, 43, and 47 in the third α-helix of the homeodomain interact with nucleotide residues in the major groove of DNA and determine, at least in part, the nucleotide sequence recognized (29–31). Since NK-1, *lab*, *Abd-B*, and *Dll* genes each contain an intron between the codons for the 44th and 45th homeobox amino acid residues, the part of each gene that is thought to encode the DNA recognition site of the corre-

```
ACG GCC CAT GCC CTA CAC AAC AAC AAT AAT AAT ACG ACA AAC AAC -160
Thr Ala His Ala Leu His Asn Asn Asn Asn Asn Thr Thr Asn Asn  -54

AAT AAC CAC AGC CTG AAG GCC GAG GGG ATC AAC GGA GCA GGC AGT -115
Asn Asn His Ser Leu Lys Ala Glu Gly Ile Asn Gly Ala Gly Ser  -39

GGT CAC GAC GAT AGC CTC AAC GAA GAT GGC ATC GAG GAG GAT ATC  -70
Gly His Asp Asp Ser Leu Asn Glu Asp Gly Ile Glu Glu Asp Ile  -24

GAC GAC GTG GAC GAC GCC GAC GGC AGT GGC GGC GGG GAT GCA AAT  -25
Asp Asp Val Asp Asp Ala Asp Gly Ser Gly Gly Gly Asp Ala Asn   -9
BamHI
                                 -1 +1
GGA TCC GAC GGT CTG CCA AAT AAG AAA CGG AAG CGA CGA GTC CTG   21
Gly Ser Asp Gly Leu Pro Asn Lys Lys Arg Lys Arg Arg Val Leu    7

TTC ACC AAG GCG CAA ACA TAT GAG CTG GAA CGT CGG TTT CGA CAA   66
Phe Thr Lys Ala Gln Thr Tyr Glu Leu Glu Arg Arg Phe Arg Gln   22

CAA CGT TAC TTG AGT GCC CCG GAA CGC GAG CAC CTG GCC AGT TTG  111
Gln Arg Tyr Leu Ser Ala Pro Glu Arg Glu His Leu Ala Ser Leu   37

ATC CGC CTG ACG CCG ACC CAG GTG AAG ATC TGG TTT CAA AAC CAT  156
Ile Arg Leu Thr Pro Thr Gln Val Lys Ile Trp Phe Gln Asn His   52

CGC TAC AAG ACG AAG CGG GCG CAA AAC GAG AAG GGC TAC GAG GGT  201
Arg Tyr Lys Thr Lys Arg Ala Gln Asn Glu Lys Gly Tyr Glu Gly   67

CAT CCT GGT CTA CTG CAC GGC CAT GCC ACC CAT CCG CAT CAC CCC  246
His Pro Gly Leu Leu His Gly His Ala Thr His Pro His His Pro   82

AGT GCC CTG CCA TCG CCC GTC GGG TAG CCGTTCCAGTTCTGGTGAGGAAC  291
Ser Ala Leu Pro Ser Pro Val Gly ***                            90

GGAAAGCCCTGCTTGGGCGATAGTTCCAAACTGGGAGCCGACTGCGTCTCCGTGTCATC  341

AGCCACCGCCACCGCCATGCAGAATGCCGCCGCCCATCACTTGGTTGCCCTAAATGGAG  400
                      Pst I
CGGCCGCCTATCAACATGCCGCTGCAG                                  440
```

FIG. 3. Nucleotide sequence and deduced amino acid sequence of the homeobox region of NK-2 genomic DNA. The homeobox domain is enclosed in a large box. Repetitive Asn residues are enclosed in a box. The acidic amino acids enclosed in boxes before the homeodomain comprise an acidic region of NK-2 protein.

```
Pst I
CTG CAG TAT TAT GCG GCG GCG ATG GAC AAC AAT AAC CAC CAT CAC -316
Leu Gln Tyr Tyr Ala Ala Ala Met Asp Asn Asn Asn His His His  -66

CAG GCA ACG GGC ACA TCG AAC TCC AGT GCC GCC GAC TAC ATG CAG -271
Gln Ala Thr Gly Thr Ser Asn Ser Ser Ala Ala Asp Tyr Met Gln  -51
                        BamHI
CGC AAA TTG GCC TAT TTT GGA TCC ACC CTC GCT GCT CCT TTG GAC -226
Arg Lys Leu Ala Tyr Phe Gly Ser Thr Leu Ala Ala Pro Leu Asp  -51

ATG AGA CGC TGC ACC AGC AAC GAT TCC G GTAAGTAACTGCACGAAATTA -177
Met Arg Arg Cys Thr Ser Asn Asp Ser A                        -26

ACGCCATTCAGGCTCTAATGGACTCTGAAAAGAACGCTACTTATTCATTGGCCTTTTGT -118

ATAGGATGTATGCTAACTTTTGGTAATTTTCCCTTTACAG AC TGC GAC TCA CCA  -6
                                         sp Cys Asp Ser Pro  -

CCG CCA TTG AGC AGT TCC CCC TCG GAG TCG CCG CTA TCC CAC GAC  -1
Pro Pro Leu Ser Ser Ser Pro Ser Glu Ser Pro Leu Ser His Asp  -
                     -1 +1
GGC AGT GGA TTG AGC CGC AAG AAG CGG TCG CGT GCC GCC TTC AGC   27
Gly Ser Gly Leu Ser Arg Lys Lys Arg Ser Arg Ala Ala Phe Ser    9

CAC GCC CAG GTC TTC GAG TTG GAG CGC CGC TTT GCC CAA CAG CGC   72
His Ala Gln Val Phe Glu Leu Glu Arg Arg Phe Ala Gln Gln Arg   24

TAC TTG TCC GGT CCG GAA CGC AGC GAG ATG GCC AAG AGC CTG CGC  117
Tyr Leu Ser Gly Pro Glu Arg Ser Glu Met Ala Lys Ser Leu Arg   39

CTG ACG GAG ACC CAG GTG AAG ATC TGG TTC CAA AAC CGC CGC TAC  16
Leu Thr Glu Thr Gln Val Lys Ile Trp Phe Gln Asn Arg Arg Tyr   5

AAG ACC AAG CGC AAG CAG ATC CAG CAG CAC GAG GCC GCC CTT TTG  2
Lys Thr Lys Arg Lys Gln Ile Gln Gln His Glu Ala Ala Leu Leu

GGT GCC AGC AAG AGG GTT CCC GTC CAA GTC TTG GTG CGA GAG GAT  25
Gly Ala Ser Lys Arg Val Pro Val Gln Val Leu Val Arg Glu Asp   84

GGC AGC ACC ACC TAC GCT CAC ATG GCT GCT CCC GGT GCT GGA CAC  29
Gly Ser Thr Thr Tyr Ala His Met Ala Ala Pro Gly Ala Gly His   99
                                                    Pst I
GGC CTC GAT CCC GCC CTG ATC AAC ATC TAC CGC CAT CAG CTG CAG  34
Gly Leu Asp Pro Ala Leu Ile Asn Ile Tyr Arg His Gln Leu Gln  11
```

FIG. 4. Nucleotide sequence and deduced amino acid sequence of the homeobox region of the NK-3 gene. The homeobox is enclosed in a box. Arrowheads represent exon–intron junctions.

sponding protein is interrupted by an intron. Further work is needed to determine whether the specificity of DNA recognition by NK-1 protein is altered by alternative splicing.

Amino acid replacements that alter the 42nd or 43rd amino acid residues in the homeobox are of special interest since they may determine part of the nucleotide sequence that is recognized by the homeobox protein. The 42nd homeobox amino acid residues of NK-2 and NK-4 are proline and alanine, respectively. The unspliced form of *Saccharomyces cerevisiae* mating-type factor a-1 has proline at this site (29); however, neither proline nor alanine has been found at this site in any metazoan homeobox protein. A proline residue would not be expected to be part of an α-helix, unlike alanine or glutamic acid residues, which promote α-helix formation. The 43rd homeobox amino acid residue of NK-1, NK-2, NK-3, and NK-4 is threonine; however, the only other homeobox proteins that contain threonine at this site are msh, Dll, lab, and ro in *Drosophila* and Hox 1.6 and Hox 7.1 in the mouse.

The amino acid sequences of most or all of these homeobox domains share other unusual features [for example, see alanine (11th amino acid residue), lysine or arginine (19th residue), glutamic acid (30th residue), tyrosine (54th residue), and serine or threonine (56th residue)]. The presence of the same or similar unusual amino acid replacements in most or all of these homeodomains provides additional evidence that the newly discovered homeobox genes are related to one another.

The combined use of probes 121 and 125, which correspond to the overlapping hexapeptides shown at the top of Fig. 7, should detect only those homeobox genes that encode the amino acid sequence Gln-Val-Lys-Ile-Trp-Phe-Gln-Asn as residues 44–51 of the third α-helix of the homeodomain, which is part of the putative nucleotide sequence recognition site of homeobox proteins. Only *zen-1*, *zen-2*, *lab*, and *cad*, in addition to the homeobox genes described in this report would be expected to give positive autoradiographic signals with both 121 and 125 probes. It is uncertain whether *msh*, *Dll*, and *Abd-B* genes would give positive signals with 121 and 125 probes because both *Dll* and *Abd-B* genes contain introns

```
TTTGAATATGATCCTAAACTAGTGCCTAGTCCCTTAACGAGTTATAACTATTATAGTTA -150

CAACTATTATATTCTACTATATTCTCGATATTGGCTATCTTTCAGAG GAT AAC AGC  -94
                                                Asp Asn Ser  -32

CAG GTG ACC TCC TCG CGT TCC GAG CTG CGA AAA AAC AGC ATC AGT  -49
Gln Val Thr Ser Ser Arg Ser Glu Leu Arg Lys Asn Ser Ile Ser  -17

GGG AAC AGC AAT CCG GGG AGC AAC AGT GGT TCC ACC AAG CCC CGG   -4
Gly Asn Ser Asn Pro Gly Ser Asn Ser Gly Ser Thr Lys Pro Arg   -2
 -1  +1
ATG AAG CGA AAG CCT CGC GTG CTC TTT TCC CAG GCA CAG GTC CTG   42
Met Lys Arg Lys Pro Arg Val Leu Phe Ser Gln Ala Gln Val Leu   14

GAG CTG GAG TGT CGC TTT CGA CTC AAA AAG TAT CTG ACG GGT GCG   87
Glu Leu Glu Cys Arg Phe Arg Leu Lys Lys Tyr Leu Thr Gly Ala   29
                            HindIII
GAG CGC GAG ATA ATC GCG CAA AAG CTT AAC CTG TCG GCC ACC CAA  132
Glu Arg Glu Ile Ile Ala Gln Lys Leu Asn Leu Ser Ala Thr Gln   44

GTG AAG ATT TGG TTC CAG AAT CGG CGC TAC AAA TCG AAA CGT GGC  177
Val Lys Ile Trp Phe Gln Asn Arg Arg Tyr Lys Ser Lys Arg Gly   59

GAC ATC GAC TGC GAG GGC ATC GCC AAG CAT CTG AAG TTG AAG TCC  222
Asp Ile Asp Cys Glu Gly Ile Ala Lys His Leu Lys Leu Lys Ser   74

GAG CCC CTG GAC TCG CCC ACT TCT CTG CCC CCG CCG ATT CCC AAC  267
Glu Pro Leu Asp Ser Pro Thr Ser Leu Pro Pro Pro Ile Pro Asn   89

CAC GTG ATG TGG CCC CCA ACC ATG CAG CAA TCG CAG CAG CAG CAG  312
His Val Met Trp Pro Pro Thr Met Gln Gln Ser Gln Gln Gln Gln  104

CAG CAT CAT GCA CAG CAG CAA CAG ATG CAG CAC ATG TAG TGGACAT  358
Gln His His Ala Gln Gln Gln Gln Met Gln His Met ***          116

TTGGCAGGACGAAGGATTCGAGTCTCTAATTTATTGCAGTTTCAAGAAGAATACATGTT  417

TTAGTCTAAGGAAACGCTCATAGTTCTAGTTCTTTGTTT                       456
```

FIG. 5. Nucleotide sequence and deduced amino acid sequence of the homeobox region of the NK-4 gene. The homeobox is enclosed in a box. A CAX repeat is underlined, which encodes repetitive Gln residues. Arrowhead represents a potential splice acceptor site.

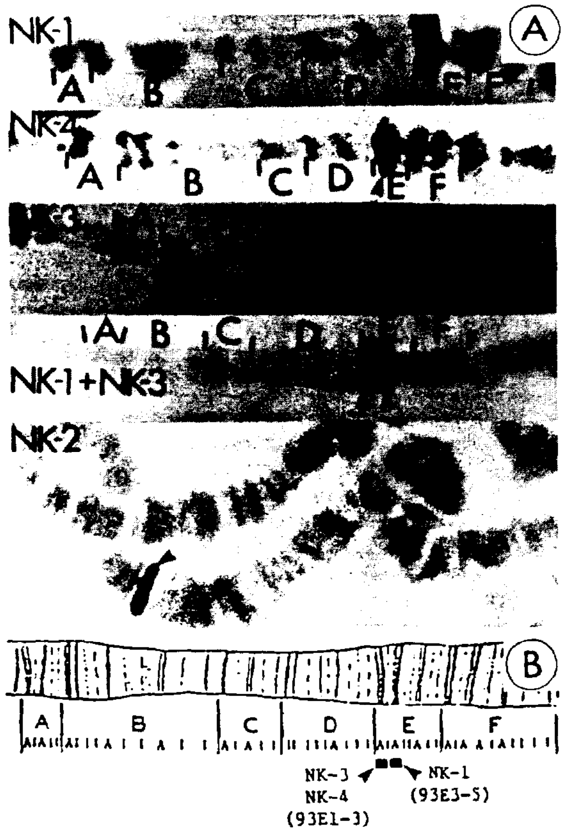

FIG. 6. (*A*) *In situ* hybridization of genomic DNA probes for NK-1 (first panel), NK-4 (second panel), NK-3 (third panel), NK-1 and NK-3 (fourth panel), and NK-2 (fifth panel) to *Drosophila* polytene chromosomes. A–F and vertical markers represent chromosomal band subdivisions in the 93 A–F region of the right arm of the third chromosome. The DNA probes hybridize to the following locations: NK-1, 93E3–5; NK-4, 93E1–3; NK-3, 93E1–3; NK-1 and NK-3, 93E1–5; NK-2, 1C1–5. Arrowheads indicate labeled chromosomal bands. (*B*) The approximate locations of the NK-3, NK-4, and NK-1 genes are indicated on Bridges' revised map of chromosomal bands (24).

of unknown sequence between codons for the 44th and 45th homeobox amino acid residues, and it is not known whether the *msh* gene contains an intron at this site. Probe 125 hybridizes to the NK-1 gene because the 3′-terminal nucleotide sequences of both exon 1 and intron 2 are CAG. Both *zen-1* and *zen-2* genomic DNA were detected and cloned with probes 121 and 125 in addition to NK-1, NK-2, NK-3, and NK-4, but *lab* and *cad* genes were not detected.

The *Drosophila* genome has been screened many times with DNA fragments containing homeobox sequences as probes; however, NK-1, NK-2, NK-3, and NK-4 homeodomains may not have been detected because their overall homology to other *Drosophila* homeodomains is relatively low. The use of oligodeoxynucleotide probes that correspond to other amino acid sequences in homeobox proteins should provide a means of detecting additional sets of homeobox genes that have some structural features in common.

1. Gehring, W. J. (1987) *Science* **236**, 1245–1252.
2. Nüsslein-Volhard, C., Frohnhöfer, H. G. & Lehmann, R. (1987) *Science* **238**, 1675–1681.
3. Scott, M. P. & Carrol, B. C. (1987) *Cell* **51**, 689–698.
4. Herr, W., Sturm, R. A., Clerc, R. G., Corcoran, L. M., Baltimore, D., Sharp, P. A., Ingraham, H. A., Rosenfeld, M. G., Finney, M., Ruvkun, G. & Horvitz, H. R. (1988) *Genes Dev.* **2**, 1513–1516.
5. Doe, C. Q. & Scott, M. P. (1988) *Trends Neurosci.* **11**, 101–106.
6. Doe, C. Q., Hiromi, Y., Gehring, W. J. & Goodman, C. S. (1988) *Science* **239**, 170–175.
7. Blochlinger, K., Bodmer, R., Jack, J., Jan, L. Y. & Jan, Y. N. *Nature (London)* **333**, 629–635.

Helices: 1 (residues 10–21), 2 (residues 28–38), 3 (residues 42–52)

Probes: #127 **EFN** ENR; #126 **ELEKEF**; **QVKIWF** #125; **KIWFQN** #121

	1	10 … 21		28 … 38		42 … 52	61	NK-1	NK-3	NK-4	NK-2	
NK-1	PRRARTAFT	**YEQLVSLENKFK**	TTRYLS	**VCERLNLALSL**	SLT	**ETQVKIWFQNR**	RTKWKKQNP	100	**54**	44	49	
NK-3	KKRSRAAFS	**HAQVFELERRFA**	QQRYLS	**GPERSEMAKSL**	RLT	**ETQVKIWFQNR**	RYKTKRKQI	**54**	100	**59**	**66**	
NK-4	KRKPRVLFS	**QAQVLELECRFR**	LKKYLT	**GAEREIIAQKL**	NLS	**ATQVKIWFQNR**	RYKSKRGDI	44	**59**	100	**59**	
NK-2	KRKRRVLFT	**KAQTYELERRFR**	QQRYLS	**APEREHLASLI**	RLT	**PTQVKIWFQNH**	RYKTKRAQN	49	**66**	**59**	100	
msh	NRKPRTPFT	**TQQLLSLEKKFR**	EKQYLS	**IAERAEFSSSL**	RLT	**ETQVKIWFQNR**	RAKAKRLQE	**57**	**54**	**54**	**54**	[29]
Dll	MRKPRTIYS	**SLQLQQLNRRFQ**	RTQYLA	**LPERAELAASL**	GLT	**QTQVKIWFQNR**	RSKYKKMMK	53	51	48	48	[22]
lab	NNSGRTNFT	**NKQLTELEKEFH**	FNRYLT	**RARRIEIANTL**	QLN	**ETQVKIWFQNR**	RMKQKKRVK	51	48	48	43	[19, 20]
zen-1	LKRSRTAFT	**SVQLVELENEFK**	SNMYLY	**RTRRIEIAQRL**	SLC	**ERQVKIWFQNR**	RMKFKKDIQ	**57**	51	44	39	[15]
zen-2	SKRSRTAFS	**SLQLIELEREFH**	LNKYLA	**RTRRIEISQRL**	ALT	**ERQVKIWFQNR**	RMKLKKSTN	49	**54**	48	41	[15]
bcd	PRRTRTTFT	**SSQIAELEQHFL**	QGRYLT	**APRLADLSAKL**	ALG	**TAQVKIWFKNR**	RRRHKIQSD	46	41	39	43	[26]
Dfd	PKRQRTAYT	**RHQILELEKEFH**	YNRYLT	**RRRRIEIAHTL**	VLS	**ERQIKIWFQNR**	RMKWKKDNK	53	48	44	38	[32]
Scr	TKRQRTSYT	**RYQTLELEKEFH**	FNRYLT	**RRRRIEIAHAL**	CLT	**ERQIKIWFQNR**	RMKWKKEHK	49	48	44	41	[33]
ftz	SKRTRQTYT	**RYQTLELEKEFH**	FNRYIT	**RRRRIDIANAL**	SLS	**ERQIKIWFQNR**	RMKSKKDRT	44	43	44	38	[30]
Antp	RKRGRQTYT	**RYQTLELEKEFH**	FNRYLT	**RRRRIEIAHAL**	CLT	**ERQIKIWFQNR**	RMKWKKENK	49	48	43	41	[34, 35]
Ubx	RRRGRQTYT	**RYQTLELEKEFH**	TNHYLT	**RRRRIEMAHAL**	CLT	**ERQIKIWFQNR**	RMKLKKEIQ	48	46	43	41	[36, 37]
abd-A	RRRGRQTYT	**RFQTLELEKEFH**	FNHYLT	**RRRRIEIAHAL**	CLT	**ERQIKIWFQNR**	RMKLKKELR	49	46	44	41	[38]
Abd-B	VRKKRKPYS	**KFQTLELEKEFL**	FNAYVS	**KQKRWELARNL**	QLT	**ERQVKIWFQNR**	RMKNKKNSQ	46	46	44	46	[21]
en	EKRPRTAFS	**SEQLARLKREFN**	ENRYLT	**ERRRQQLSSEL**	GLN	**EAQIKIWFQNK**	RAKIKKSTG	46	43	38	34	[16]
inv	DKRPRTAFS	**GTQLARLKHEFN**	ENRYLT	**EKRRQQLSGEL**	GLN	**EAQIKIWFQNK**	RAKLKKSSG	49	46	39	39	[39]
BSH4	QRRSRTTFT	**AEQLEALERAFS**	RTQYPD	**VYTREELAQTT**	ALT	**EARIQVWFSNR**	RARLRKHSG	44	41	33	34	[40]
BSH9	QRRSRTTFS	**NDQIDALERIFA**	RTQYPD	**VYTREELAQST**	GLT	**EARVQVWFSNR**	RARLRKQLN	48	44	38	38	[40]
prd	QRRCRTTFS	**ASQLDELERAFE**	RTQYPD	**IYTREELAQRT**	NLT	**EARIQVWFSNR**	RARLRKQHT	44	34	30	30	[25]
ro	QRRQRTTFS	**TEQTLRLEVEFH**	RNEYIS	**RSRRFELAETL**	RLT	**ETQIKIWFQNR**	RAKDKRIEK	51	51	44	46	[8, 9]
cad	KDKYRVVYT	**DFQRLELEKEYC**	TSRYIT	**IRRKSELAQTL**	SLS	**ERQVKIWFQNR**	RAKERTSNK	43	41	44	39	[41]
H2.0	RSWSRAVFS	**NLQRKGLEIQFQ**	QQKYIT	**KPDRRKLAARL**	NLT	**DAQVKVWFQNR**	RMKWRHTRE	39	46	41	39	[42]
eve	VRRYRTAFT	**RDQLGRLEKEFY**	KENYVS	**RPRRCELAAQL**	NLP	**ESTIKVWFQNR**	RMKDKRQRI	41	36	34	33	[43, 44]
cut	SKKQRVLFS	**EEQKEALRLAFA**	LDPYPN	**VGTIEFLANEL**	GLA	**TRTITNWFHNH**	RMRLKQQVP	31	28	33	31	[7]
	OO ● OO	● ●O O	●	O OO O	●	O OOOO●●O●O	● O OO					

FIG. 7. Comparison of the amino acid sequences (single-letter code) of NK-1, NK-2, NK-3, and NK-4 homeoboxes with known *Drosophila* homeoboxes. The positions of three α-helices identified by Otting *et al.* (31) in a peptide containing the *Antp* homeobox are indicated by boxes 1–3 above the sequences and the corresponding amino acid residues are shown in boldface type. Oligodeoxynucleotide probes 121–127 correspond to the amino acid sequences shown at the top. Arrowhead above the NK-1 sequence represents the location of introns in NK-1, *lab*, *Dll*, and *Abd-B* genes. Chromosomal clusters of homeobox genes are NK-1, NK-3, and NK-4; *lab–Antp*; *Ubx–Abd-B*; *en* and *inv*; and *BSH-4* and *BSH-9* (*gsb*). Symbols at the bottom of the table represent the following for *Drosophila* homeoboxes: ●, invariant amino acid residues; ○, strongly conserved amino acid residues. The percentage homology of the amino acid sequences of NK-1, NK-2, NK-3, or NK-4 homeoboxes compared with other *Drosophila* homeoboxes are shown on the right. Values represent percentage of amino acid residues that are identical in each pair of homeoboxes compared; 100% corresponds to 61 amino acid residues. Numbers in brackets are references.

8. Saint, R., Kalionis, B., Lockett, T. J. & Elizur, A. (1988) *Nature (London)* **334**, 151–154.
9. Tomlinson, A., Kimmel, B. E. & Rubin, G. M. (1988) *Cell* **55**, 771–784.
10. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab., Cold Spring Harbor, NY).
11. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. & Efstratiadis, A. (1978) *Cell* **15**, 687–701.
12. Wood, W. I., Gitschier, J., Laskey, L. A. & Lawn, R. M. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1585–1588.
13. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
14. Pardue, M. L. (1986) in *Drosophila: A Practical Approach*, ed. Roberts, D. B. (IRL, Oxford), pp. 111–137.
15. Rushlow, C., Doyle, H., Hoey, T. & Levin, M. (1987) *Genes Dev.* **1**, 1268–1279.
16. Poole, S. J., Kauvar, L. M., Dress, B. & Kornberg, T. (1985) *Cell* **40**, 37–43.
17. Biggin, M. D., Bickel, S., Benson, M., Pirrotta, V. & Tjian, R. (1988) *Cell* **53**, 713–722.
18. Müller, M., Affolter, M., Leupin, W., Otting, G., Würthrich, K. & Gehring, W. J. (1988) *EMBO J.* **7**, 4299–4304.
19. Mldozik, M., Fjose, A. & Gehring, W. J. (1988) *EMBO J.* **7**, 2569–2578.
20. Hoey, T., Doyle, H. J., Harding, K., Wedeen, C. & Levin, M. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 4809–4813.
21. DeLorenzi, M., Ali, N., Saari, G., Henry, C., Wilcox, M. & Bienz, M. (1988) *EMBO J.* **7**, 3223–3231.
22. Cohen, S. M., Brönner, G., Küttner, F., Jürgens, G. & Jäckel, H. (1989) *Nature (London)* **338**, 432–434.
23. Sunkel, C. E. & Whittle, J. R. S. (1987) *Wilhelm Roux's Arch. Dev. Biol.* **196**, 124–132.
24. Bridges, P. N. (1941) *J. Hered.* **32**, 299–300.
25. Frigerio, G., Burri, M., Bopp, D., Baumgartner, S. & Noll, M. (1986) *Cell* **47**, 735–746.
26. Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M. & Nüsslein-Volhard, C. (1988) *EMBO J.* **7**, 1749–1756.
27. Zink, B. & Paro, R. (1989) *Nature (London)* **337**, 468–471.
28. Demerec, M., Kaufman, B. P., Fano, U., Sutton, E. & Sansome, E. R. (1942) *Carnegie Inst. Washington Yearb.* **41**, 190–218.
29. Gehring, W. J. (1987) in *Molecular Approaches to Developmental Biology*, eds. Firtel, R. A. & Davidson, E. H. (Liss, New York), pp. 115–129.
30. Laughon, A. & Scott, M. P. (1984) *Nature (London)* **310**, 25–31.
31. Otting, G., Qian, Y., Müller, M., Affolter, M., Gehring, W. & Wüthrich, K. (1988) *EMBO J.* **7**, 4305–4309.
32. Regulski, M., McGinnis, N., Chadwick, R. & McGinnis, W. (1987) *EMBO J.* **6**, 767–777.
33. LeMotte, P. K., Kuroiwa, A., Fessler, L. I. & Gehring, W. J. (1989) *EMBO J.* **8**, 219–227.
34. Schneuwly, S., Kuroiwa, A., Baumgartner, P. & Gehring, W. J. (1986) *EMBO J.* **5**, 733–739.
35. Laughon, A., Boulet, A. M., Bermingham, J. R., Jr., Laymon, R. A. & Scott, M. P. (1986) *Mol. Cell. Biol.* **6**, 4676–4689.
36. Weinzierl, R., Axton, J. M., Ghysen, A. & Akam, M. (1987) *Genes Dev.* **1**, 386–397.
37. Kornfeld, K., Saint, R. B., Beachy, P. A., Harte, P. J., Peattie, D. A. & Hogness, D. S. (1989) *Genes Dev.* **3**, 243–258.
38. Akam, M., Dawson, I. & Tear, G. (1988) *Development* **104**, Suppl. 123–133.
39. Coleman, K. G., Poole, S. J., Weir, M. P., Soeller, W. C. & Kornberg, T. (1987) *Genes Dev.* **1**, 19–28.
40. Baumgartner, S., Bopp, D., Burri, M. & Noll, M. (1987) *Genes Dev.* **1**, 1247–1267.
41. Mlodzik, M. & Gehring, W. J. (1987) *Cell* **48**, 465–478.
42. Barad, M., Jack, T., Chadwick, R. & McGinnis, W. (1988) *EMBO J.* **7**, 2151–2161.
43. Macdonald, P. M., Ingham, P. & Struhl, G. (1986) *Cell* **47**, 721–734.
44. Frash, M., Hoey, T., Rushlow, C., Doyle, H. & Levin, M. (1986) *EMBO J.* **6**, 749–759.